2

GIBBS SAMPLING FOR MARGINAL POSTERIOR EXPECTATIONS

ALAN E. GELFAND
AND
ADRIAN F. M. SMITH

TECHNICAL REPORT NO. 448

NOVEMBER 19, 1991

PREPARED UNDER CONTRACT
N00014-89-J-1627 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

2

GIBBS SAMPLING FOR MARGINAL POSTERIOR EXPECTATIONS

Alan E. Gelfand
and
Adrian F. M. Smith

Technical Report No. 448

November 19, 1991

Prepared Under Contract
N00014-89-J-1627 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

GIBBS SAMPLING FOR MARGINAL POSTERIOR EXPECTATIONS

Alan E. Gelfand
Department of Statistics
University of Connecticut
Storrs, Connecticut, 06269-3120, USA

Adrian F.M. Smith
Department of Mathematics
Nottingham University
Nottingham, England, NG7 2RD

ABSTRACT

In earlier work (Gelfand and Smith, 1990 and Gelfand et al, 1989) a sampling based approach using the Gibbs sampler was offered as a means for developing marginal posterior densities for a wide range of Bayesian problems several of which were previously inaccessible. Our purpose here is two-fold. First we flesh out the implementation of this approach for calculation of arbitrary expectations of interest. Secondly we offer comparison with perhaps the most prominent approach for calculating posterior expectations, analytic approximation involving application of the LaPlace method. Several illustrative examples are discussed as well. Clear advantages for the sampling based approach emerge.

Key Words: Bayesian inference, marginal posterior expectations, Gibbs sampler, LaPlace method, Aggregated Multinomial Model, Variance Components Model, Normal-Linear Hierarchical Model.

GIBBS SAMPLING FOR MARGINAL POSTERIOR EXPECTATIONS

Alan E. Gelfand
Department of Statistics
University of Connecticut
Storrs, Connecticut, 06269–3120, USA

Adrian F. M. Smith
Department of Mathematics
Nottingham University
Nottingham, England, NG7 2RD

1. Introduction

An important issue in Bayesian inference is the calculation of marginal posterior expectations. Let $[\underset{\sim}{Y}|\underset{\sim}{\Theta}]$ denote the likelihood where $\underset{\sim}{Y}$ is the observed data and $\underset{\sim}{\Theta}$ is a vector of unobservable random parameters. Specifying a prior distribution for $\underset{\sim}{\Theta}$ yields the marginal posterior distribution for $\underset{\sim}{\Theta}$ denoted by $[\underset{\sim}{\Theta}|\underset{\sim}{Y}]$. Interest typically centers on calculating the marginal posterior distribution itself along with associated expectations such as $E(\Theta_i|\underset{\sim}{Y})$, $E(\Theta_i^2|\underset{\sim}{Y})$, $E(\Theta_i\Theta_j|\underset{\sim}{Y})$, $E(1_{a,b}(\Theta_i|\underset{\sim}{Y}))$ where $1_{a,b}$ denotes the indicator function of the interval (a,b). In many applications functions of $\underset{\sim}{\Theta}$, $g(\underset{\sim}{\Theta})$ are of interest e.g., Θ_i-Θ_j, Θ_i/Θ_j, max Θ_i, $E(W(Y)|\mathbf{\Theta})$ (for an appropriate W) where again the marginal posterior distribution and associated expectations would be sought.

The technical problems encountered in attempting to carry out the required numerical integrations to obtain these distributions and expectations have long served as an impediment to the wider application of the Bayesian framework to real data.

In two recent papers (Gelfand and Smith, 1990, Gelfand et al 1989) the Gibbs sampler was discussed in the context of calculation of marginal posterior densities for a wide range of Bayesian problems several of which were previously inaccessible. The Gibbs sampler offers conceptual simplicity and straightforward implementation avoiding sophisticated numerical or analytic approximation expertise and associated specialist software (as in e.g. Naylor and Smith 1982, 1988, Smith et al 1985, 1987, Shaw 1988, Geweke 1988). The Gibbs sampler was proposed for and has been implemented for very high dimensional problems (Geman and Geman. 1984) e.g. 10^5 or more variables. By comparison even a complicated Bayesian application would usually have low dimension. Thus in our applications this sampler has proved very efficient converging remarkably quickly. Indeed we need not compromise high dimensional integration by replacing integration with estimation as is often done in practice.

The purpose of this note is to show that not only does the Gibbs sampler enable calculation of marginal posterior densities but it also enables routine calculation of expectations such as those described above.

Since our emphasis here is on expectations, in the next section we review existing approaches both analytic and numerical. In the third section we briefly review the Gibbs sampler followed by the development of its application to calculating expectations. In Section 4 we provide a variety of illustrative applications. We conclude in Section 5 with a brief summary.

2. Existing Approaches for Posterior Expectations

With regard to performing the required numerical integrations directly there have been several recent advances (see e.g. Naylor and Smith 1982. 1988. Smith et al 1985. 1987). As noted in the introduction, implementation requires both numerical expertise and highly sophisticated software. These approaches employ quadrature methods, are typically most successful in integrating functions which are of the form "polynomial × normal density" and will usually not accommodate more than six dimensions. (Of course insightful manipulation of joint distributions can frequently reduce higher dimensional problems to forms requiring at most six dimensional integration).

An alternative direction for numerical integration has been through Monte Carlo methods as in the work of Stewart (1983, 1984), Van Dijk and Kloek (1980, 1984) and Shaw (1988) perhaps in its most refined form using importance sampling and variance reduction techniques in recent work of Geweke (1988, 1989). Recently Rubin (1987, 1988) has proposed a sampling/importance resampling algorithm which in the context of hierarchical models affords the advantage of simpler specification of the importance sampling density. Again these approaches encounter difficulty with high dimensionality. Moreover our empirical experience (see, for example, Gelfand and Smith, 1990) suggests that the "learning" which is inherent in an iterative algorithm such as the Gibbs sampler makes such an approach more efficient in terms of amount of random generation than one–off (non iterative) Monte Carlo methods.

Substantial effort has been devoted to the development of analytic approximations for calculating expectations. Most of this work involves application of Laplace's method (see e.g. DeBruijn, 1961). Writing these expectations as a ratio of integrals and approximating numerator and denominator separately Tierney and Kadane (1986) obtain a second order approximation for the expectation of positive functions. Other second order expansions (Hartigan 1965, Johnson 1970, Lindley 1961, 1980) require computation of more derivatives of the log likelihood function. Extension of the Tierney–Kadane approach to nonpositive functions is discussed in Tierney, Kass and Kadane (1989a). Kass and Steffey (1988) study these approximations in the context of conditionally independent hierarchical models. Density approximations based on the Laplace method appear in Tierney and Kadane (1986) with extensions in Tierney, Kass and Kadane (1989b). A related approximation appears in Hsu, Leonard and Tsui (1987). Work of Achcar and Smith (1989) shows that performance of the Laplace method is often very sensitive to parametrization. Morris (1988) offers expansions based on Pearson family kernels rather than the normal kernels used in the Laplace method.

Apart from the sometimes severe sensitivity to parametrization (see Section 4) these approaches require at least one and usually two function maximizations which often demand expertise and are at best cumbersome. In addition, these functions must be assumed twice differentiable. Moreover a separate function maximization is required for each expectation sought. These problems are magnified with increasing dimensionality.

Since the Laplace method has received much attention of late its performance along with that of the Gibbs sampler will be investigated in Section 4. We thus briefly review the method encouraging the reader to consult Kass, Tierney and Kadane (1988) for a fuller account.

Consider a set of random variables $\underset{\sim}{U} = (U_1,\cdots.U_k)$. Our interest is in calculating $E\,f(\underset{\sim}{U})$. Suppose $f>0$ and suppose the joint density of $U_1,\cdots,U_k$ is only known modulo normalizing constant i.e. is proportional to $g(U_1,\cdots,U_k)$. Then $E(f) =$

$$\frac{\int f\cdot g}{\int g} = \frac{\int e^{-\ell^*}}{\int e^{-\ell}} \approx \left[\frac{|\Sigma^*|}{|\Sigma|}\right]^{\frac{1}{2}} e^{\ell(\hat{\underset{\sim}{U}})-\ell^*(\hat{\underset{\sim}{U}}^*)} \qquad (1)$$

where $\ell = -\log g$, $\ell^* = \ell-\log f$, $\hat{\underset{\sim}{U}}$ is the mode of ℓ, $\hat{\underset{\sim}{U}}^*$ is the mode of ℓ^*, Σ and Σ^* are minus the inverse Hessians of ℓ and ℓ^* evaluated at $\hat{\underset{\sim}{U}}$ and $\hat{\underset{\sim}{U}}^*$ respectively. The form (1) first appeared in Tierney and Kadane (1986) who noted that when $\log g = 0(n)$ this approximation is then accurate to order n^{-2}.

Tierney, Kass and Kadane (1989a) suggest extending (1) to handle nonpositive f by approximating the moment generating function of f, $E(\exp(sf))$, and then differentiating at $s=0$. Alternatively we may add a large constant c to f such that $c+f>0$, apply (1) to $c+f$, and then subtract c from the resulting approximation. If c needs to be infinite we regain the approximation based on the moment generating function.

Let $h_s \equiv E(f(\underset{\sim}{U})\,|\,U_r,r\neq s)$. If the joint density of the $U_r,r\neq s$ is proportional to a known function say $g_s(U_r,\ r=1,\cdots k,\ r\neq s)$ then, since $E(f)=E(h_s)$, $E(f) =$

$$\frac{\int h_s g_s}{\int g_s} = \frac{\int e^{-\ell_s^*}}{\int e^{-\ell_s}} \approx \left[\frac{|\Sigma_s^*|}{|\Sigma_s|}\right]^{\frac{1}{2}} e^{\ell_s-\ell_s^*} \qquad (2)$$

where ℓ_s, ℓ_s^* etc. are defined analogous to (1). We see that (2) replaces the approximation of a ratio of k–fold integrals with the approximation of a ratio of (k-1)–fold integrals and thus should provide a better approximation than (1). Hierarchical Bayes models provide a class of probability structures where approximation (2) can usually be carried out (see Kass & Steffey, 1989).

3. Gibbs Sampling

3.1 Review of the Gibbs Sampler

In the sequel we assume the existence of densities with respect to either Lebesque or counting measure as appropriate. Densities will be denoted, generically, by square brackets so that joint, conditional and marginal forms appear, respectively, as $[U,V]$ $[U|V]$ and $[V]$. The usual marginalization by integration procedure will be denoted by forms such as

$$[U] = \int [U|V]^*[V]$$

We shall require that our collections of random variables are such that specification of all full conditional distributions uniquely determines the full joint density (see Besag, 1974). More precisely, for such a collection of random variables $U_1, U_2, \cdots, U_k$, the joint density, $[U_1, U_2, \cdots, U_k]$, is assumed uniquely determined by $U_s|U_r,\ r \neq s]$, $s=1,2,\cdots,k$. Our interest is in the marginal distributions, $[U_s,]$, $s=1,2,\cdots k$.

An algorithm for extracting marginal distributions from the full conditional distribution was formally introduced as the Gibbs sampler in Geman and Geman (1984). The algorithm requires all the full conditional distributions to be "available" for sampling, where "available" is taken to mean that, for example, U_s can be generated straightforwardly and efficiently given specified values of the conditioning variables, U_r, $r \neq s$. We return to this matter at the end of this section.

Gibbs sampling is a Markovian updating scheme which proceeds as follows. Given an arbitrary starting set of values $U_1^{(0)},\cdots,U_k^{(0)}$, we draw $U_1^{(1)}$ from $[U_1|U_2^{(0)},\cdots,U_k^{(0)}]$, then $U_2^{(1)}$ from $[U_2|U_1^{(1)}, U_3^{(0)},\cdots,U_k^{(0)}]\cdots$ and so on up to $U_k^{(1)}$ from $[U_k|U_1^{(1)},\cdots,U_{k-1}^{(1)}]$ to complete one iteration of the scheme. After t such iterations we would arrive at $(U_1^{(t)},\cdots,U_k^{(t)})$. Geman and Geman show under mild conditions that $(U_1^{(t)},\cdots,U_k^{(t)}) \xrightarrow{d} (U_1,\cdots,U_k) \sim [U_1, U_2,\cdots U_k]$ as $t\to\infty$. Hence for t large enough $U_s^{(t)}$ for example will be regarded as a simulated observation from $[U_s]$. Replication of the process m times yields m iid k–tuples $(U_{1j}^{(t)},\cdots,U_{kj}^{(t)})$ $j=1,\cdots,m$. Diagnostics to assess convergence are critical. Some tentative discussion appears in Gelfand et al (1989) but a fuller account is deferred to a future paper. Note that sample size at say the t^th^ iteration may be increased from m to any specified size by sampling with replacement from the vectors $(U_{1j}^{(t)},\cdots,U_{kj}^{(t)})$, $j=1,\cdots m$.

Application in the Bayesian framework takes the U_s to be unobservable representing either parameters or missing data. All distributions will be viewed as conditional on the observed data. $[U_1,\cdots U_k]$ becomes the joint posterior density whose form is therefore known modulo normalizing constant. There is no question as to whether the full conditional densities uniquely determine the joint density: the full conditional densities will have been obtained from the form of the joint density functions. Functions. $f(U_1,\cdots,U_k)$, whose density and expectation we seek will arise as interesting functions of the parameters.

We conclude this section with a remark regarding the required generation from full conditional distributions. In a hierarchical Bayesian model the full conditional distributions take reduced forms (see Gelfand and Smith, 1990, Section 3.2). In particular if conjugacy is assumed for the distribution at a given pair of adjacent stages say $i,i+1$, in the hierarchy then the full posterior at stage $i+1$ is an updated version of the prior at stage $i+1$ whence sampling is usually straightforward. If conjugacy is not assumed then the full posterior for any parameter will still be known modulo normalizing constant. In this case more sophisticated random generation using for example the ratio of uniforms method (see e.g. Devroye. 1986) still enables sampling from the full posterior.

3.2 Calculating Expectations

Since the iid vectors $(U_{1j}^{(t)},\cdots,U_{kj}^{(t)})$ are approximately distributed as $[U_1,\cdots,U_k]$ we can in principle create sample based estimates of the marginal densities of any subset of the U's or indeed of any transformation of the U's using kernel density estimators (see e.g. Silverman, 1986). In the same spirit an obvious sample based estimator of the expected value of an integrable function $f(U_1,\cdots,U_k)$ takes the form

$$\sum_{j=1}^{m} f(U_{1j}^{(t)},\cdots,U_{kj}^{(t)})/m \tag{3}$$

We now show that in all such enterprise it behooves us to take advantage, when possible, of availability of the full conditional densities to improve our estimation. As in the previous section, for any s, $E\, f(U_1,\cdots,U_k) = E(h_s)$ where $h_s = E(f|U_r,\ r\neq s)$. But $\text{var}(f) \geq \text{var}(h_s)$. Hence the estimator

$$\sum_{j=1}^{m} h_s(U_{rj}^{(t)},\ r\neq s)/m \tag{4}$$

is better than (3) in terms of mean squared error. In fact the "Rao–Blackwellized" estimator (4) is better than (3) under more general loss functions (see e.g. Ferguson 1967, p. 121). We typically take advantage of this Rao–Blackwellization by using m smaller than would be required under (3). We note that (3) and (4) are sampling analogues of (1) and (2) respectively and that (2) may be viewed as a Rao–Blackwellized version of (1). We also note that the sampling based approach does not require smoothness or differentiability for f and h_s allowing application to for example $f=\max U_i$.

With regard to density estimation again if U_s appears as an argument of f the conditional density $[f \mid U_r, r \neq s]$ can be obtained by univariate transformation from $[U_s \mid U_r, r \neq s]$. The resulting Rao–Blackwellized sample–based density estimate of f would be

$$[\hat{f}]_s = \sum_{j=1}^{m} [f \mid U_{rj}^{(t)}, r \neq s]/m \qquad (5)$$

Note that we need not modify the Gibbs sampler so as to sample f's in order to calculate E(f) or to estimate [f]. Note further that the forms (3), (4) and (5) are invariant under 1–1 transformation of the individual U_i. Hence, in practical calculation of expectations using the Gibbs sampler, transformation of the parameters is not an issue as it is for the Laplace method.

By using the full conditional distribution of each variable which actually appears as an argument of f, the form (4) may, in principle, be used to obtain several estimators which improve upon (3). Similarly, several density estimators of the form (5) can be obtained. While adaptive combination of such estimators might be attempted (say by approximating the variance of each estimator through the delta–method or through a sample reuse method) we suggest the simple unweighted average. The variance components example in Section 4.2 provides an illustration.

A related point involves the use of reduced conditional distributions and expectations when available to improve estimation. We make the argument in its simplest form taking k=3. To calculate $E(U_1)$ suppose both $h(U_2,U_3)=E(U_1 \mid U_2,U_3)$ and $W(U_2)=E(U_1 \mid U_2)$ are known in closed form. Since $E(h \mid U_2)=EW$, Rao–Blackwellization shows that $m^{-1} \sum_{j=1}^{m} W(U_{2j}^{(t)})$ is better than $m^{-1} \sum_{j=1}^{m} h(U_{2j}^{(t)},U_{3j}^{(t)})$. This point concurs with our intuition regarding Monte Carlo integration for a fixed number of independent points, m – we can approximate a single integral better than a double integral. This situation typically arises in the context of missing data. For instance in the aggregated multinomial example of section 4.1 the reduced conditional distributions $[\Theta | Y, Z]$ and $[\eta | Y, Z]$ are immediately unscaled Beta distributions.

3.3 A Conditional Gibbs Sampler

The Gibbs sampler can be extended to yield samples from arbitrary conditional distributions as follows. Suppose we fix a subset of the U_r's, without loss of generality say $U_{k'+1}, \cdots, U_k$ at $u_{k'+1}, \cdots, u_k$ respectively. In implementing the Gibbs sampler suppose a complete iteration is achieved by updating only $U_1, \cdots, U_{k'}$. That is, starting with $(U_1^{(0)}, \cdots, U_{k'}^{(0)}, U_{k'+1}^{(0)}, \cdots U_k^{(0)})$ we draw $U_1^{(1)} \sim [U_1 | U_2 = U_2^{(0)}, \cdot, U_{k'} = U_{k'}^{(0)}, U_{k'+1} = u_{k'+1}, \cdots U_k = u_k] \cdots, U_{k'}^{(1)} \sim [U_{k'} | U_1^{(1)}, \cdots U_{k'-1}^{(1)}, U_{k'+1}, = u_{k'+1}, \cdots U_k = u_k]$. As the number of such iterations, $t \to \infty$, $(U_1^{(t)}, \cdots U_{k'}^{(t)}) \xrightarrow{d} (U_1, \cdots U_{k'}) \sim [U_1, \cdots, U_{k'} | U_{k'+1} = u_{k'+1}, \cdots U_k = u_k]$. Repeating this process m times yields $(U_{1_j}^{(t)}, \cdots, U_{k'j}^{(t)})$, $j = 1, \cdots m$ an iid sample having approximately this conditional distribution.

Estimation would then proceed as in Section 3.2. For instance a Rao–Blackwellized density estimate for $[U_2 | U_1 = u_1]$ would be developed holding U_1 fixed at u_1 in the Gibbs sampler and would take the form

$$\widehat{[U_2 | U_1} = u_1] = \sum_{j=1}^{m} [U_2 | U_1 = u_1, U_r = U_{rj}^{(t)}, r > 2]/m \qquad (6)$$

A Rao–Blackwellized estimate for $E(U_2 | U_1 = u_1)$ would take the form

$$\sum_{j=1}^{m} E(U_2 | U_1 = u_1, U_r = u_{rj}^{(t)}, r > 2)/m \qquad (7)$$

4. Illustrative Examples

In this section we illustrate the methodology of Section 3 in the context of three examples chosen to reflect different types of "awkward" structure where the Gibbs sampler provides an easily implemented solution.

4.1. An Aggregated Multinomial Model

Gelfand and Smith (1990) discuss a fictitious data set where some observations are not assigned to individual cells but to aggregates of cells (see e.g. Hartley, 1958; Dempster, Laird & Rubin 1977, Tanner & Wong, 1987). In fact suppose the data $Y=(Y_1,\cdots,Y_5) = (14,1,1,1,5)$ are available as a sample from the multinomial distribution

$$\text{Mult}\,(22; \tfrac{1}{4}\,\theta + \tfrac{1}{8}, \tfrac{1}{4}\,\theta, \tfrac{1}{4}\,\eta, \tfrac{1}{4}\,\eta + \tfrac{3}{8}, \tfrac{1}{2}(1-\theta-\eta))$$

and that, as a prior for (θ,η), we take Dirichlet (1,1,1). By considering instead a "split cell" multinomial of the form

$X=(X_1,\cdots,X_7) \sim \text{Mult}(22; \tfrac{1}{4}\,\theta, \tfrac{1}{8}, \tfrac{1}{4}\,\theta, \tfrac{1}{4}\,\eta, \tfrac{1}{4}\,\eta, \tfrac{3}{8}, \tfrac{1}{2}(1-\theta-\eta))$

we view X_1, X_5 as missing data and construct a Gibbs sampler involving θ, η and $Z=(X_1, X_5)$. The required full conditional distributions are

$$[\theta\,|\,Y,\eta,Z] = (1-\eta)\text{ Be}\,(X_1 + Y_2+1, Y_5+1) \tag{8}$$

$$[\eta\,|\,Y,\Theta,Z] = (1-\theta)\text{ Be}\,(Y_3 + X_5+1, Y_5+1) \tag{9}$$

and

$$[Z\,|\,Y,\theta,\eta] = [X_1, X_5\,|\,Y,\theta,\eta] = \text{Bi}\,(Y_1, 2\theta(1+2\theta)^{-1}) \cdot \text{Bi}\,(Y_4, 2\eta(3+2\eta)^{-1})$$

Using "exact" numerical methods (Naylor and Smith, 1982) we obtain $E[\theta|Y] = .5199$, $E[\eta|Y] = .1232$. The Tierney–Kadane approximation (1) on the (θ,η) scale yields $E(\theta|Y) \approx .5175$ and $E(\eta|Y) \approx .0882$ resulting in relative errors of .5% and

28.4% respectively. Under an alternative parameterization (for example the logit is considered in Achcar and Smith, 1989) we might fare better with $E(\eta|Y)$.

For the sampling based approaches in addition we considered $E(\theta^2|Y)$, $E(\eta^2|Y)$, $\mathrm{var}(\theta|Y)$ and $\mathrm{var}(\eta|Y)$, a total of six expectations. Simulation is required to study the performance of these approaches. We obtained 5000 repetitions of the Gibbs sampler where for each repetition we took only 10 iterations and set m to be only 20. For each repetition we calculated estimates of the six expectations using the form (3) and using the form (4), for the later employing (8) or (9) as appropriate. For instance, for $E(\theta|Y)$ the estimator of the form (3) is $\sum_{j=1}^{20} \theta_j^{(10)}/20$ while the estimator of the form (4) employing (8) is $\sum_{j=1}^{20} d_j/20$ where $d_j = (1 - \eta_j^{(10)})(X_{ij}^{(10)} + Y_2 + 1)(X_{ij}^{(10)} + Y_2 + Y_5 + 2)^{-1}$. For $\mathrm{var}(\theta|Y)$ the estimator of the form (3) is $\sum_{j=1}^{20} (\theta_j^{(10)} - \bar{\theta}^{(10)})^2/20$ where $\bar{\theta}^{(10)}$ is the average of the $\theta_j^{(10)}$. But since

$$\mathrm{var}(\theta|Y) = E(\mathrm{var}(\theta|Y,\eta,Z)) + \mathrm{var}\, E(\theta|Y,\eta,Z) \qquad (10)$$

we can use $\mathrm{var}(\theta|Y,\eta,Z)$ from (8) to obtain an estimator of the form (4) for the first term on the right hand side of (10). For the second term on the right hand side we use the estimator $\Sigma(d_j - \bar{d})^2/19$ where $\bar{d} = \Sigma d_j/20$.

For both of the estimators of each of the six expectations over the 5000 replications we calculated the average, the average bias, the variance and the mean square error. In all cases bias was inconsequential in the sense that mean square error agrees with variance to at least three significant places. Hence in Table 1 we present, for each expectation the "exact" value, the average value using the form (3) estimator with associated standard error in parentheses and the average value using form (4) estimator with associated standard error in parentheses. The performance of the Gibbs sampler is remarkably good especially given such small t and m. In practice, by the last iteration, it would typically be $\geq$1000 yielding SE's at most one–tenth of those shown. As it is, for, say $E[\eta|Y]$ using the estimator of form (4) virtually all estimates fell in the interval (.1228. .1236). Note also the benefit of "Rao–Blackwellizing". In all cases the standard error of the form (2) estimators is at most 58% of the corresponding form (1) estimator. Thus the MSE's of the Rao Blackwellized estimators are at most 35% their non–conditional counterparts.

[INSERT TABLE 1 HERE]

4.2 A Variance Components Model

Box and Tiao (1973, Chapter 5) present a set of dyestuff data wherein five samples of six randomly chosen batches of raw material were taken and a single laboratory determination of product yield was made for each of the resulting 30 samples. If Y_{ij} denotes the yield of the j_{th} sample from the i_{th} batch, $i=i,\cdots,6$, $j=1,\cdots,5$ then we define a variance components model for Y_{ij} by $Y_{ij} = \theta_i + \epsilon_{ij}$ where, assuming conditional independence throughout, $[\Theta_i \mid \mu,\sigma_\Theta^2] = N(\mu,\sigma_\Theta^2)$, $[\epsilon_{ij} \mid \sigma_e^2] = N(0,\sigma_e^2)$. Given Θ_i and σ_e^2, for the i_{th} batch Y_i and S_i^2 are sufficient and thus we summarize the data as

Batch	1	2	3	4	5	6
$\bar{Y}_i$	1505	1528	1564	1498	1600	1470
S_i	63.05	33.28	37.98	68.70	50.00	31.02

The usual ANOVA estimate of σ_Θ^2 is 1764.05, of σ_e^2 is 2451.25. Let Y denote $\{Y_{ij}\}$, let $\underset{\sim}{\Theta}=(\Theta_1,\cdots,\Theta_2)$ and assume that μ,σ_Θ^2 and σ_e^2 are independent with priors $[\mu] = N(\mu_0,\sigma_0^2)^2$, $[\sigma_0^2] = IG(a_1,b_1)$ and $[\sigma_1^2] = IG(a_2,b_2)$ where IG denotes the inverse Gamma distribution and $\mu_0,\sigma_0^2,a_1,b_1,a_2,b_2$ are assumed known. For illustrative purposes we take the rather vague specifications $\mu_0=0$, $\sigma_0^2=10^{12}$, $a_1=0$, $b_1=1$, $a_2=0, b_2=1$. The Gibbs sampler involves $\sigma_\Theta^2,\sigma_e^2$, μ and $\underset{\sim}{\Theta}$. The required full conditional distributions are

$$[\sigma_{\theta}^{2}|\underset{-}{Y},\mu,\underset{-}{\Theta},\sigma_{1}^{2}] = \text{IG}\ (3.5,\ 1000 + (.5)\Sigma(\Theta_{i}-\mu)^{2}) \qquad (11)$$

$$[\sigma_{e}^{2}|\underset{-}{Y},\mu,\underset{-}{\Theta},\sigma_{\theta}^{2}] = \text{IG}\ (15,\ 4\Sigma S_{i}^{2} + 5\Sigma(\bar{Y}_{i}-\Theta_{i})^{2}) \qquad (12)$$

$$[\mu|\underset{-}{Y},\underset{-}{\Theta},\sigma_{\theta}^{2},\sigma_{1}^{2}] = N(\Sigma\,\Theta_{i/6},\ \sigma_{\theta}^{2}/6)$$

$$[\underset{-}{\Theta}|\underset{-}{Y},\mu,\sigma_{\theta}^{2},\sigma_{e}^{2}] = N((\sigma_{e}^{2}+5\sigma_{\theta}^{2})^{-1}(5\sigma_{\theta}^{2}\bar{Y} + \sigma_{e}^{2}1),\ (\sigma_{e}^{2}+5\sigma_{\theta}^{2})^{-1}I)$$

where $Y'=(\bar{Y}_1,\cdots,\bar{Y}_6)$, 1 is a 6×1 column of 1's and I is a 6×6 identity matrix.

Interest typically focuses on the distributions $[\sigma_{\theta}^{2}|Y]$, $[\sigma_{e}^{2}|Y]$ and $[\sigma_{\theta}^{2},\sigma_{e}^{2}|Y]$. The marginal distribution distributions are straightforwardly handled using estimates of the form (5) and have been created for a different and in fact more difficult data set in Gelfand et al (1989). With regard to expectations we consider $E(\sigma_{\theta}^{2}|Y)$, $E(\sigma_{e}^{2}|Y)$ and $E(\sigma_{\theta}^{2}/\sigma_{e}^{2}|Y)$. Using "exact" numerical methods (Naylor and Smith, 1982) these expectations are 2496.6, 2879.4, and 1.019 respectively. Using the Tierney–Kadane approximations (1) on the $(\sigma_{e}^{2},\sigma_{\theta}^{2})$ scale yields very poor estimates, $E(\sigma_{\theta}^{2}|Y) = 5378.5$, $E(\sigma_{e}^{2}|Y) = 6393.9$, and $E(\sigma_{\theta}^{2}/\sigma_{e}^{2}) = 1.044$. Achcar and Smith (1989) report a less severe but similar finding and recommend a log transformation of the variance components. (In fact, the exact answers above were obtained working with log scales). Unfortunately on the log scale the results are not much better: $E(\sigma_{\theta}^{2}|Y) = 1679.7$, $E(\sigma_{e}^{2}|Y) = 3669.0$ and $E(\sigma_{\theta}^{2}/\sigma_{e}^{2}|Y) = 2.959$.

In addressing the performance of the Gibbs sampler we focused on $E(\sigma_{\theta}^{2}/\sigma_{e}^{2}|Y)$. Using a simulation based on 5000 repetitions with t=60, m=50 at each repetition, we examined four estimators:

- estimator 1 is of the form (3) i.e. $\left(\sum_{j=1}^{50} \sigma_{\theta j}^{2(30)} \big/ \sigma_{ej}^{2(30)}\right)/50$

– estimator 2 is of the form (4) using (11) i.e. $\left(\sum_{j=1}^{50} \frac{1000+\Sigma(\Theta_{ij}^{(30)}-\mu_j^{(30)})^2}{2.5\,\sigma_{1j}^{2(30)}}\right)/50$

– estimator 3 is of the form (4) using (12) i.e. $\left(\sum_{j=1}^{50} \frac{15\,\sigma_{\Theta j}^{2(30)}}{4\Sigma 5_i^2 + 5\Sigma\bar{Y}_i-\Theta_{ij}^{(30)}}\right)/50$

– estimator 4 is the unweighted average of estimators 2 and 3.

The brief Table 2 reports the exact value along with the average value and standard error for each of the four estimators. The performance of the Gibbs sampler is quite good especially since m is still much smaller than we would use in practice. While, after the fact, estimator 3 performs better than estimator 2 (one had to be better than the other!), in the absence of such knowledge estimator 4 seems a satisfactory choice. As a matter of record this simulation yielded for $E(\sigma_\theta^2 | Y)$ an average value of 2500.1 with standard error 322.8, for $E(\sigma_e^2 | Y)$ an average value of 2877.8 with standard error 85.7. Again, in practice, a much larger m would be used. As noted in Section 3 transformations need not be considered.

INSERT TABLE 2 HERE

Note that if interest was in the intra-class correlation coefficient, $\sigma_\theta^2/(\sigma_\theta^2+\sigma_e^2)$ we may straightforwardly obtain Rao–Blackwellized density estimates for $[\sigma_\theta^2/(\sigma_\theta^2+\sigma_e^2) | Y]$ using (5). However in calculating $E(\sigma_\theta^2/(\sigma_\theta^2+\sigma_e^2) | Y)$ only an estimator of the form (3) is readily available.

4.3 A Normal–linear Hierarchical Model

A widely used version of the normal–linear hierarchical model introduced by Lindley and Smith (1972) takes the following general form. Data on the i$^{\text{th}}$ of k individuals is modeled by $[Y_i \mid \boldsymbol{\theta}_i, \sigma^2] = N(X_i \boldsymbol{\theta}_i, \sigma^2 I_{n_i})$, with the individual parameters $\boldsymbol{\theta}_i$ themselves modeled by $[\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, \Sigma] = N(\boldsymbol{\mu}, \Sigma)$, conditional independence for $i=1, \cdots, k$ being assumed throughout. The Bayesian modeling is then completed by assuming, independently, that $[\boldsymbol{\mu}] = N(\boldsymbol{\eta}, C)$, $[\Sigma^{1}] = W((\rho R)^{-1}, \rho)$ and $[\sigma^2] = IG(\frac{1}{2}\nu_0, \frac{1}{2}\nu_0 \tau_0^{2})$. Here, W and IG denote Wishart and inverse–gamma distributions, respectively. Typical applications of such a model are to population random effects studies (see, for example, Racine–Poon and Smith, 1989).

In such cases, the $[Y_i \mid \boldsymbol{\theta}_i, \sigma^2]$ describe n_i individual measurements (e.g. growth over time/response to dose), the k regression coefficients $\boldsymbol{\theta}_i$ characterize the individual growth or response pattern and the $[\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, \Sigma_i]$ reflect the fact that the k individuals are a random sample from a population. A variety of Bayesian and empirical Bayesian procedures have been proposed for making inferences about $\boldsymbol{\theta}_i$, $\boldsymbol{\mu}$, Σ, σ^2, but exact calculations have hitherto proved infeasible, largely due to the presence of the unknown population covariance matrix Σ. Discussion and illustration of inferences and predictions for such hierarchical models is given in Gelfand et. al. (1989). Here we simply indicate how the Gibbs sampler provides a straightforward and easily implemented approach to the general problem of estimating Σ

Defining $Y = (Y_1, \cdots, Y_k)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k)$, $\bar{\boldsymbol{\theta}} = k^{-1} \sum_{i=1}^{k} \boldsymbol{\theta}_i$, $n = \sum_{i=1}^{k} n_i$, $D_i = \sigma^{-2} X_i^T X_i + \Sigma^{1}$, $V = (k\Sigma^{1} + C^{-1})^{-1}$, the Gibbs sampler is easily seen to take the form (suppressing explicit dependence on $\boldsymbol{\eta}, C, \rho, R, \nu_0, \tau_0^2$)

$$
\begin{aligned}
[\boldsymbol{\theta}_i \mid Y, \boldsymbol{\mu}, \Sigma^{1}, \sigma^2] &= N(D_i(\sigma^{-2} X_i^T Y_i + \Sigma^{1} \boldsymbol{\mu}), D_i) \ (i=1, \cdots, k) \\
[\boldsymbol{\mu} \mid Y, \boldsymbol{\theta}, \Sigma^{1}, \sigma^2] &= N(V(k\Sigma^{1} \bar{\boldsymbol{\theta}} + C^{-1} \boldsymbol{\eta}), V) \\
[\Sigma^{1} \mid Y, \boldsymbol{\theta}, \boldsymbol{\mu}, \sigma^2] &= W([\sum_i (\boldsymbol{\theta}_i - \boldsymbol{\mu})(\boldsymbol{\theta}_i - \boldsymbol{\mu})^T + \rho R]^{-1}, k+\rho) \qquad (13) \\
[\sigma^2 \mid Y, \boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma^{1}] &= IG(\tfrac{1}{2}(n+v_0), \tfrac{1}{2}[\sum_i (Y_i - X_i \boldsymbol{\theta}_i)^T (Y_i - X_i \boldsymbol{\theta}_i) + \nu_0 \tau_0^2).
\end{aligned}
$$

Simulation from the N and IG distributions is immediate; simulation from the W distribution is straightforwardly accomplished using the algorithm of Odell and Feiveson (1966). Estimation of Σ (via Σ^{1}) using the form (4) is directly achieved using the mean of

(13). As an indication of the computing effort required, for an example with $k=30$, $n_i=5$ and $X_i\theta_i$ having the form of a straight–line growth curve (so that $\theta, \mu, \Sigma, \sigma^2$ involves a total of 66 parameters), satisfactory convergence (checked by empirical Q-Q plots for the eigenvalues of Σ^{-1}) was obtained at $t=25$ with $m=100$.

5. Discussion

Our general discussion and the three illustrative examples reveals the ease with which posterior expectations can be calculated using the Gibbs sampler, even with "awkward" model/data combinations. In particular, we note the striking advantage of the approach in overcoming problems due to: awkward posterior surfaces, otherwise requiring subtle and sophisticated numerical or analytic approximation techniques; functions which may not be continuous much less differentiable; dimensionality problems arising from highly parameterized models; intractable distributions arising from missing data; "difficult" objects, such as covariance matrices, whose component parameters implicitly involve complicated constraints.

Most attractively, we need not redo the analysis for each desired expectation. All such expectations are readily estimated from the final sample.

References

Achcar, J.A. and Smith, A.F.M. (1989). Aspects of reparametrization in approximate Bayesian inference. In *Essays in Honor of George A. Barnard*, J. Hodges editor; North–Holland.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society B*, 36, 192–326.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian inference in statistical analysis.* Addison–Wesley, Reading, MA.

DeBruijn, N.G. (1961). *Asymptotic Methods In Analysis.* Amsterdam, North Holland.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the E–M Algorithm (with Discussion). *Journal of the Royal Statistical Society, B*, 39, 1–38.

Devroye, L. (1986). *Non–uniform random variate generation.* Springer–Verlag, NY.

Ferguson, T. (1967). Mathematical Statistics: *A Decision Theoretic Approach.* Academic Press, N.Y.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities, *Journal of American Statistical Association* (to appear).

Gelfand, A.E., Hills, S.E., Racine–Poon, A., Smith, A.F.M. (1989) Illustration of Bayesian Inference In Normal Data Models Using Gibbs Sampling. Department of Mathematics Technical Report, University of Nottingham.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geweke, J. (1988). Antithetic acceleration of Monte Carlo integration in Bayesian inference. *Journal of Econometrics*, 38, 73–90.

Geweke, J. (1989). Bayesian Inference In Econometric Models Using Monte Carlo Integration. *Econometrica* (to appear).

Hartigan, J.A. (1965). The asymptotically unbiased prior distribution. *Annals of Mathematical Statistics*, 36, 1137–1152.

Hartley, H.O. (1958). Maximum Likelihood Estimation from Incomplete Data. *Biometrics*, 14, 174–194.

Hsu, J.S., Leonard, T. and Tsui, K. (1987). Bayesian and likelihood marginal inference with applications to contingency table analysis. Technical Report, Department of Statistics, University of Wisconsin.

Johnson, R.A. (1970). Asymptotic Expansions for Posterior Distributions. *Annals of Mathematical Statistics*. 41, 1266–1272.

Kass, R., Tierney, L. and Kadane, J.B. (1988). Asymptotics in Bayesian Computation in *Bayesian Statistics*, 3. J.M. Bernardo et al, editors, Oxford University Press. 261–278.

Kass, R.E. and Steffey, D.L. (1989). Approximate Inference in Conditionally Independent Hierarchical Models. (Parametric Empirical Bayes Models) *Journal of American Statistical Association*, 84, 717–726.

Lindley, D.V. (1961). "The Use of Prior Probability Distributions In Statistical Inference and Decisions." Proc. 4th Berkeley Symp. 1, 453–468.

Lindley, D.V. (1980). "Approximate Bayesian Methods" in *Bayesian Statistics*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith, University Press, Valencia, Spain.

Morris, C. "Approximating Posterior Distributions and Posterior Moments" In: *Bayesian Statistics* 3, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Eds. Oxford University Press, p. 327–344.

Naylor, J.C. and Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distribution. *Applied Statistics*, 31, 214–225.

Naylor, J.C. and Smith, A.F.M. (1988). Econometric illustrations of novel numerical integration strategies for Bayesian inference. *Journal of Econometrics*. 38, 103–126.

Odell, P.L. and Feiveson, A.H. (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association*, 61, 198–203.

Racine–Poon, A. and Smith, A.F.M. (1989). Population Models. To appear in *Statistical methodology in the Pharmaceutical Sciences* (Ed. D. Berry). M. Dekker, Inc., NY.

Shaw, J.E.H. (1988). A quasirandom approach to integration in Bayesian statistics. *Annals of Statistics*, 16, 895–914.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C. and Dransfield, M. (1985). The implementation of the Bayesian paradigm. *Communications in Statistics, Theory and Methods*, 14, 1079–1102.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C. (1987). Progress with numerical and graphical methods for Bayesian statistics. *Statistician*, 36, 75–82.

Stewart, L. (1983). Bayesian Analysis Using Monte Carlo Integration – A Powerful Methodology for Handling Some Difficult Problems. In: *Practical Bayesian Statistics*, A.P. David and A.F.M. Smith, Eds., Harlow, England, Longman.

Stewart, L. (1984). Multiparameter Parameter Bayesian Inference Using Monte Carlo Integration – Some Techniques for Bivariate Analysis. In *Bayesian Statistics*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, Eds., North Holland, Amsterdam.

Tanner, M. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Assoc.* 82, 528–50.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82–86.

Tierney, L., Kass, R.E. and Kadane, J.B. (1989a). Fully exponential LaPlace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84, 710–716.

Tierney, L., Kass, R.E. and Kadane, J.B. (1989b). Approximate Marginal Densities of Nonlinear Functions, *Biometrika*, 76, 425–436.

Van Dijk, H.K. and Kloek, T. (1980). Further experience in Bayesian Analysis Using Monte Carlo Integration, *Journal of Econometrics*, 14, p. 307–328.

Van Dijk, H.K. and Kloek, T. (1984). Experiment with some alternative for simple importance sampling in Monte Carlo integration. In *Bayesian Statistics*. 2, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith, Eds., North Holland, Amsterdam.

Table 1: Expectations for the Aggregated Multinomial Model

Expectation	Exact Value	Average Value of Form (3) Estimator (S.E.)	Average Value of Form (4) Estimator (S.E.)
$E(\eta\|Y)$	.1232	.1229 (.0178)	.1232 (.0082)
$E(\Theta^2\|Y)$	.2881	.2879 (.0302)	.2883 (.0171)
$E(\eta^2\|Y)$	$.2174\times10^{-1}$	$.2158\times10^{-1}$ ($.0639\times10^{-1}$)	$.2173\times10^{-1}$ ($.0280\times10^{-1}$)
$Var(\Theta\|Y)$	$.1776\times10^{-1}$	$.1781\times10^{-1}$ ($.0551\times10^{-1}$)	$.1774\times10^{-1}$ ($.0218\times10^{-1}$)
$Var(\eta\|Y)$	$.6561\times10^{-2}$	$.6476\times10^{-2}$ ($.2732\times10^{-2}$)	$.6556\times10^{-2}$ ($.0902\times10^{-2}$)

Table 2: The Expected Variance Ratio for the Variance Components Model

Exact	Estimator 1	Estimator2	Estimator 3	Estimator 4
Value	(SE)	(SE)	(SE)	(SE)
1.019	1.095	1.086	1.028	1.057
	(.098)	(.091)	(.022)	(.039)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
448

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Gibbs Sampling for Marginal Posterior Expectations

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Alan E. Gelfand and Adrian F. M. Smith

8. CONTRACT OR GRANT NUMBER(s)
N00014-89-J-1627

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
Stanford University
Stanford, CA 94305

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR-042-267

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Statistics & Probability Program Code 111

12. REPORT DATE
November 19, 1991

13. NUMBER OF PAGES
21

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release: distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Bayesian inference, marginal posterior expectations, Gibbs sampler, Laplace method, aggregated multinomial model, variance components model, normal-linear hierarchical model

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Please see the following page

Technical Report No. 448

Abstract

In earlier work (Gelfand and Smith, 1990 and Gelfand et al., 1989) a sampling based approach using the Gi-bs sampler was offered as a means for developing marginal posterior densities for a wide range of Bayesian problems several of which were previously inaccessible. Our purpose here is two-fold. First we flesh out the implementation of this approach for calculation of arbitrary expectations of interest. Secondly we offer comparison with perhaps the most prominant approach for calculating posterior expectations, analytic approximation involving application of the Laplace method. Several illustrative examples are discussed as well. Clear advantages for the sampling based approach emerge.